IN THE CLAIMS:

Please cancel claims 2-33 and amend the remaining claim in the application as follows:

1. (Currently Amended) A method for managing persistent connections between data processing units of a computer system, wherein a first data processing unit is connected to a second data processing unit to send requests to the second data processing unit for processing, the method comprising the steps of:

monitoring a communication delay period for requests transferred from the first data processing unit to the second data processing unit;

comparing the monitored delay period with a threshold communication delay period to determine whether the monitored communication delay period indicates a predefined performance condition; ~~and,~~

in response to determining that the monitored communication delay period indicates a predefined performance condition, adjusting the number of connections between the first and second data processing units,

wherein the comparing step comprises determining whether the monitored delay period exceeds a first threshold delay period,

wherein the step of adjusting the number of connections is responsive to determining that the monitored delay period exceeds the first threshold to establish at least one additional connection,

wherein the comparing step comprises determining whether the monitored delay period is less than a second threshold delay period,

wherein the step of adjusting the number of connections is responsive to determining that the monitored delay period is less than the second threshold to close at least one connection,

wherein the comparing step comprises determining whether the monitored delay period exceeds a first threshold delay period,

wherein the step of adjusting the number of connections is responsive to determining that the monitored delay period exceeds the first threshold to establish at least one additional connection,

wherein the monitored delay period is calculated by:

computing a difference between (1) a timestamp associated with the transfer of the request from the first data processing unit to the second data processing unit and (2) a timestamp associated with the receipt at the first data processing unit of a response to said request from the second data processing unit, and subtracting a time period measured as the time processing the request within the second data processing unit,

wherein the monitored communication delay period is averaged for a set of requests processed during a predefined time period prior to said step of comparing the monitored communication delay period with a threshold communication delay period;

wherein said method further comprises the step of calibrating the system to determine a first threshold communication delay period above which the establishment of at least one additional connection is expected to reduce the communication delay period,

wherein the step of calibrating the system to determine a first threshold comprises:

monitoring communication delay periods and corresponding request throughput information for different numbers of concurrent clients and different numbers of persistent connections;

determining a minimum number of concurrent clients at which a predefined percentage increase in throughput can be achieved by increasing the number of persistent connections by an integer value, *a*, between the first and second data processing units;

identifying, with reference to the monitored communication delay periods; and

corresponding request throughput information, a communication delay period corresponding to the determined minimum number of concurrent clients,

wherein the step of calibrating the system is performed separately for each of CPU-intensive requests and data-intensive requests, and the step of calibrating comprises the additional step of selecting a minimum communication delay period from the identified

communication delay period for CPU-intensive requests and the identified communication delay period for data-intensive request,

wherein said method further comprises the step of using said calibration of the system to determine a second threshold communication delay period below which the closing of at least one connection is not expected to significantly increase the communication delay period,

wherein the first and second threshold communication delay periods are computed as percentage differences from a selected minimum communication delay period,

wherein said method further comprises the step of performed prior to the step of establishing at least one additional connection, of checking whether the adjusted number of connections would exceed a maximum permitted number of connections,

wherein the step of establishing the at least one additional connection is performed only if the adjusted number of connections would not exceed the maximum permitted number,

wherein said method further comprises the step of monitoring communication delay periods and corresponding request throughput information for different numbers of concurrent clients and different numbers of persistent connections,

wherein said method further comprises the step of determining the maximum permitted number of connections by: identifying a maximum throughput, for different numbers of concurrent clients, from the monitored request throughput information; identifying a number of persistent connections corresponding to said identified maximum throughput for each respective number of concurrent clients; and selecting a minimum from said identified numbers of persistent connections,

wherein the first data processing unit is a front-end network gateway node of a cluster-based data processing system, and the second data processing unit is a back-end processing node of the cluster-based data processing system,

wherein said method further comprises the step of the gateway node receiving requests from a client requestor via a network, passing received requests to respective ones of a set of back-end processing nodes, receiving responses from the respective back-end processing nodes, and forwarding received responses to the client requestor; and

wherein said method further comprises the step of the back-end processing node processing requests received from the gateway node to generate responses, and forwarding the responses to the gateway node,

wherein the cluster-based data processing system comprises a plurality of back-end processing nodes, and

wherein the method is responsive to monitored communication delays for the plurality of back-end processing nodes to modify the number of persistent connections consistently for the plurality of back-end processing nodes.

2-33. (Cancelled).